

K-Nearest Neighbors (KNN)

Math 243

Michael Pearce

Outline

Last time, we. . .

- Define **Mean Squared Error (MSE)** for regression problems
- Compared models using the **Bias-Variance Trade-Off**

Today, we will. . .

- Discuss the **Bayes Classifier**
- Implement **KNN** as estimate for Bayes Classifier

The Task: Classification

Let:

- Y be a **categorical response** variable with levels $A_1, \dots, A_k$
- $X_1, \dots, X_p$ are **predictors** (either categorical or quantitative)

The Task: Classification

Let:

- Y be a **categorical response** variable with levels $A_1, \dots, A_k$
- $X_1, \dots, X_p$ are **predictors** (either categorical or quantitative)

Goal: Build a model $\hat{f}(X_1, \dots, X_p)$ such that

- $\hat{f}$ is a function of the predictors $X_1, \dots, X_p$
- $\hat{f}$ predicts the class of Y

$$\text{e.g., } \hat{f}(x_{01}, \dots, x_{0p}) = A_1$$

Q: How can we measure the accuracy of this model? Can we use MSE?

The Task: Classification

Let:

- Y be a **categorical response** variable with levels $A_1, \dots, A_k$
- $X_1, \dots, X_p$ are **predictors** (either categorical or quantitative)

Goal: Build a model $\hat{f}(X_1, \dots, X_p)$ such that

- $\hat{f}$ is a function of the predictors $X_1, \dots, X_p$
- $\hat{f}$ predicts the class of Y

$$\text{e.g., } \hat{f}(x_{01}, \dots, x_{0p}) = A_1$$

Q: How can we measure the accuracy of this model? Can we use MSE?

$$\text{Nope! } \text{MSE} = \frac{1}{n} \sum_{i=1}^n \left(\underbrace{y_i - \hat{f}(x_{01}, \dots, x_{0p})}_{\text{e.g., Green} - \text{Red} = ??} \right)^2$$

Error rate

Training data: Compute *error rate* on observations in *training* data:

$$\text{Training Error} = \frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{f}(x_i))$$

where $I(y_i \neq \hat{f}(x_i))$ is the **incorrect classification Indicator function**,

$$I(y_i \neq \hat{f}(x_i)) = \begin{cases} 1 & \text{if } y_i \neq \hat{f}(x_i) \\ 0 & \text{otherwise} \end{cases}$$

Error rate

Training data: Compute *error* rate on observations in *training* data:

$$\text{Training Error} = \frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{f}(x_i))$$

where $I(y_i \neq \hat{f}(x_i))$ is the **incorrect classification Indicator function**,

$$I(y_i \neq \hat{f}(x_i)) = \begin{cases} 1 & \text{if } y_i \neq \hat{f}(x_i) \\ 0 & \text{otherwise} \end{cases}$$

Test data: Compute average proportion of *errors* on *test* data

$$\text{Test Error} = \text{Avg. } I(y_i \neq \hat{f}(x_0))$$

with the average taken across many test observations x_0 .

Bayes Classifier: The Best Possible Model

The response, Y , likely depends on more than just $X_1, \dots, X_p$:

- Even a “perfect” model has irreducible error, ϵ

Bayes Classifier: The Best Possible Model

The response, Y , likely depends on more than just $X_1, \dots, X_p$:

- Even a “perfect” model has irreducible error, ϵ

In classification, the “perfect” model is called the **Bayes Classifier**:

$$f(x_0) = \operatorname{argmax}_{A_j} P(Y = A_j \mid X = x_0)$$

- **In words**: Given predictor value(s) x_0 , assign Y to the outcome, A_j , with highest probability.
- “Perfect” means the Bayes Classifier minimizes *test* error.

Bayes Classifier: The Best Possible Model

The response, Y , likely depends on more than just $X_1, \dots, X_p$:

- Even a “perfect” model has irreducible error, ϵ

In classification, the “perfect” model is called the **Bayes Classifier**:

$$f(x_0) = \operatorname{argmax}_{A_j} P(Y = A_j | X = x_0)$$

- **In words**: Given predictor value(s) x_0 , assign Y to the outcome, A_j , with highest probability.
- “Perfect” means the Bayes Classifier minimizes *test* error.
- We usually don’t know $P(Y = A_j | X = x_0) \implies$ can’t build this “perfect” model
- But we can **estimate** it!!

Simulation

- Suppose $Y \in \{A, B\}$ and predictors X_1, X_2 are each between 0 and 1.

Simulation

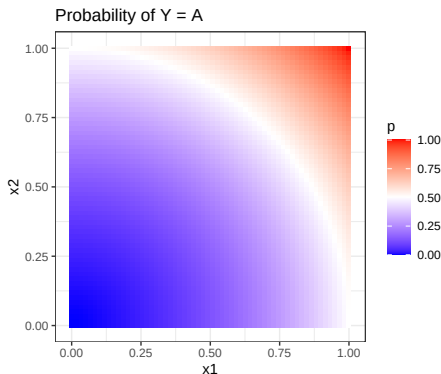
- Suppose $Y \in \{A, B\}$ and predictors X_1, X_2 are each between 0 and 1.
- Additionally, suppose true model is determined by

$$P(Y = A | x_1, x_2) = \frac{x_1^2 + x_2^2}{2}$$

Simulation

- Suppose $Y \in \{A, B\}$ and predictors X_1, X_2 are each between 0 and 1.
- Additionally, suppose true model is determined by

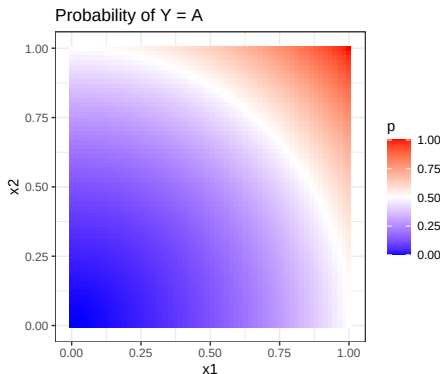
$$P(Y = A | x_1, x_2) = \frac{x_1^2 + x_2^2}{2}$$



Simulation

- Suppose $Y \in \{A, B\}$ and predictors X_1, X_2 are each between 0 and 1.
- Additionally, suppose true model is determined by

$$P(Y = A | x_1, x_2) = \frac{x_1^2 + x_2^2}{2}$$

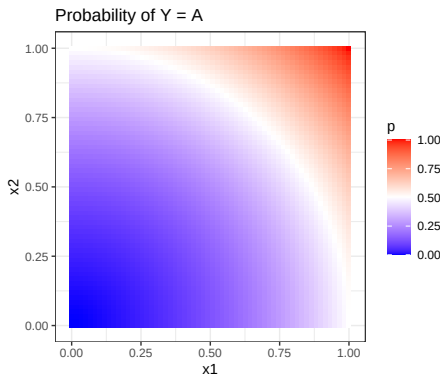


- Q: What is the Bayes Classifier, f ?

Simulation

- Suppose $Y \in \{A, B\}$ and predictors X_1, X_2 are each between 0 and 1.
- Additionally, suppose true model is determined by

$$P(Y = A | x_1, x_2) = \frac{x_1^2 + x_2^2}{2}$$

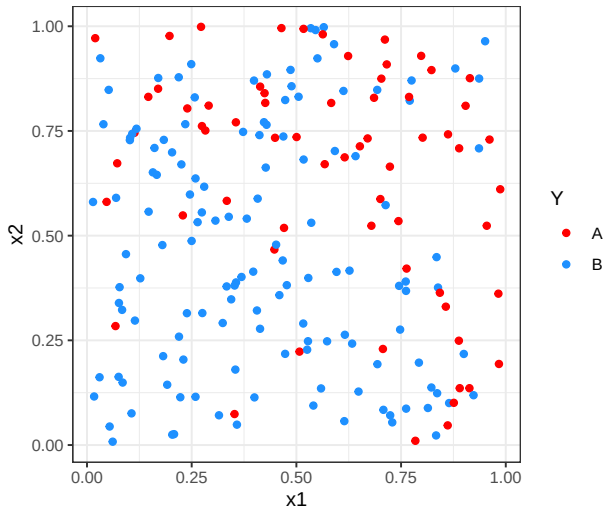


- **Q:** What is the Bayes Classifier, f ?

$$\begin{aligned} f(x_0) &= \underset{A_j}{\operatorname{argmax}} P(Y = A_j | X = x_0) \\ &= \begin{cases} A & \text{if } \frac{x_1^2 + x_2^2}{2} \geq 0.5 \\ B & \text{if } \frac{x_1^2 + x_2^2}{2} < 0.5 \end{cases} \end{aligned}$$

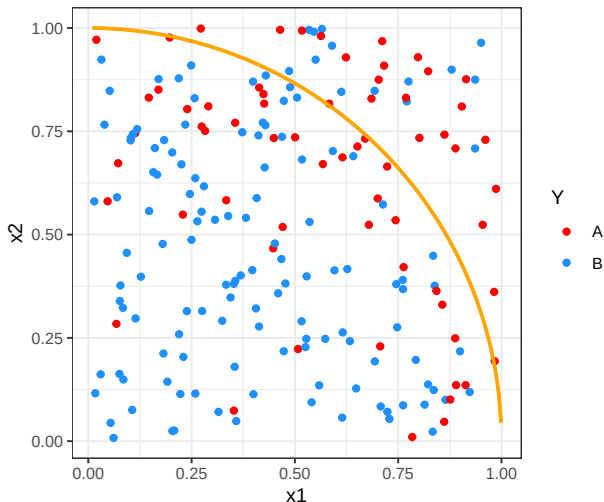
Simulate Data

Let's simulate 200 data points from this model.



The Bayes Classifier

The orange arc represents the Bayes Classifier boundary



Expected Error Rate

The Bayes Classifier produces an averaged error rate of

$$\text{Avg.} \left(1 - \max_j P(Y = A_j | X = x_0) \right)$$

where the average is over X .

Expected Error Rate

The Bayes Classifier produces an averaged error rate of

$$\text{Avg.} \left(1 - \max_j P(Y = A_j | X = x_0) \right)$$

where the average is over X .

- Q: Why does this make sense?

Expected Error Rate

The Bayes Classifier produces an averaged error rate of

$$\text{Avg.} \left(1 - \max_j P(Y = A_j | X = x_0) \right)$$

where the average is over X .

- **Q:** Why does this make sense?
- In our problem, error rate ≈ 0.274 . (Verify with calculus or large simulation)

Expected Error Rate

The Bayes Classifier produces an averaged error rate of

$$\text{Avg.} \left(1 - \max_j P(Y = A_j | X = x_0) \right)$$

where the average is over X .

- **Q:** Why does this make sense?
- In our problem, error rate ≈ 0.274 . (Verify with calculus or large simulation)

Note: This is the **theoretical** lowest error rate for this classification problem.

- Any actual model will do worse than this, on average.
- Analogous to irreducible error, ϵ , in regression problems.

Section 2

K-Nearest Neighbors (KNN)

From Bayes Classifier to KNN

In *theory*, the Bayes Classifier is our best model for classification.

From Bayes Classifier to KNN

In *theory*, the Bayes Classifier is our best model for classification.

- In *practice*, we don't know the conditional probability of Y given X , and so cannot build a Bayes Classifier.

From Bayes Classifier to KNN

In *theory*, the Bayes Classifier is our best model for classification.

- In *practice*, we don't know the conditional probability of Y given X , and so cannot build a Bayes Classifier.
- But given enough data, we can *estimate* the model

From Bayes Classifier to KNN

In *theory*, the Bayes Classifier is our best model for classification.

- In *practice*, we don't know the conditional probability of Y given X , and so cannot build a Bayes Classifier.
- But given enough data, we can *estimate* the model

One such estimate. . .

K-Nearest Neighbors (KNN): look at the K closest *training* points to test point x_0 , and choose the most common class among them

- KNN is a **non-parametric** statistical learning method

K-Nearest Neighbors (KNN), Mathematically

K-Nearest Neighbors (KNN) Given,

K (positive integer), x_0 (test observation)

K-Nearest Neighbors (KNN), Mathematically

K-Nearest Neighbors (KNN) Given,

K (positive integer), x_0 (test observation)

Let N_0 denote K nearest *training* observations to x_0 . Then, KNN

(1) Calculates

$$\hat{P}_j(x_0) = \hat{P}(Y = A_j | X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = A_j)$$

Q: In words, what does this mean?

K-Nearest Neighbors (KNN), Mathematically

K-Nearest Neighbors (KNN) Given,

K (positive integer), x_0 (test observation)

Let N_0 denote K nearest *training* observations to x_0 . Then, KNN

(1) Calculates

$$\hat{P}_j(x_0) = \hat{P}(Y = A_j | X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = A_j)$$

Q: In words, what does this mean?

- $\hat{P}_j(x_0)$ = proportion of the K nearest *training* points to x_0 that were in class A_j

K-Nearest Neighbors (KNN), Mathematically

K-Nearest Neighbors (KNN) Given,

K (positive integer), x_0 (test observation)

Let N_0 denote K nearest *training* observations to x_0 . Then, KNN

(1) Calculates

$$\hat{P}_j(x_0) = \hat{P}(Y = A_j | X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = A_j)$$

Q: In words, what does this mean?

- $\hat{P}_j(x_0)$ = proportion of the K nearest *training* points to x_0 that were in class A_j

(2) Classifies the test observation as:

$$\hat{g}(x_0) = \operatorname{argmax}_{A_j} \hat{P}_j(x_0)$$

Q: In words, what does this mean?

K-Nearest Neighbors (KNN), Mathematically

K-Nearest Neighbors (KNN) Given,

K (positive integer), x_0 (test observation)

Let N_0 denote K nearest *training* observations to x_0 . Then, KNN

(1) Calculates

$$\hat{P}_j(x_0) = \hat{P}(Y = A_j | X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = A_j)$$

Q: In words, what does this mean?

- $\hat{P}_j(x_0)$ = proportion of the K nearest *training* points to x_0 that were in class A_j

(2) Classifies the test observation as:

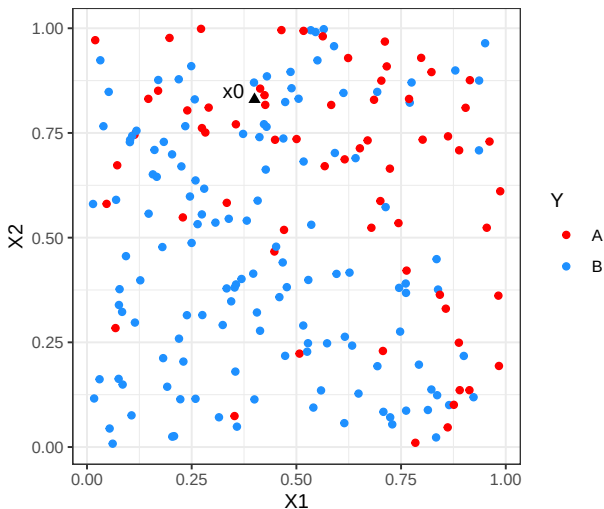
$$\hat{g}(x_0) = \operatorname{argmax}_{A_j} \hat{P}_j(x_0)$$

Q: In words, what does this mean?

- Label x_0 as the most common class (A_j) among the K nearest *training* points

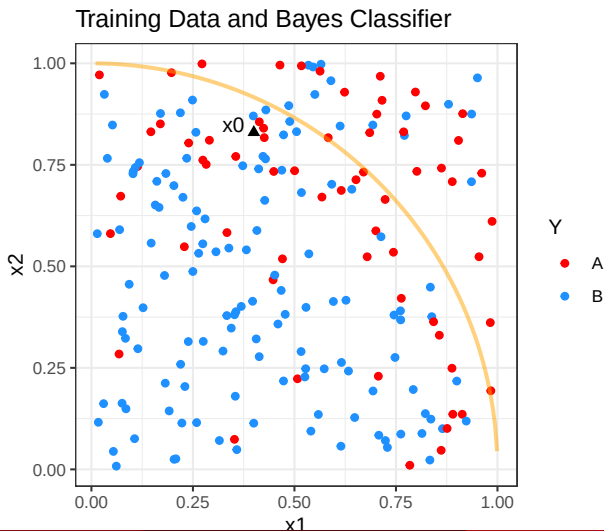
Example

Let's classify x_0 for $K = 1, 2, 5, 10, 50, 100$.



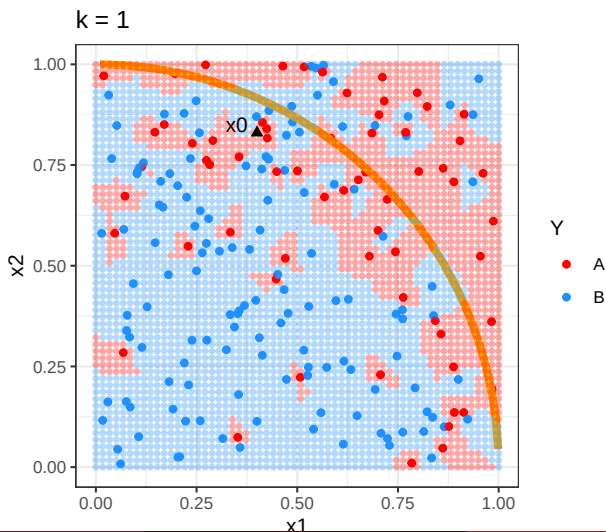
Classification Boundaries

Here are the classification boundaries for a variety of values of K .



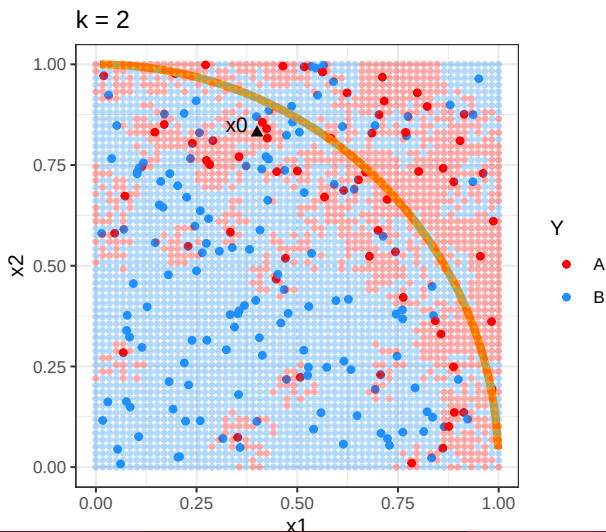
$k=1$

Here are the classification boundaries for a variety of values of K .



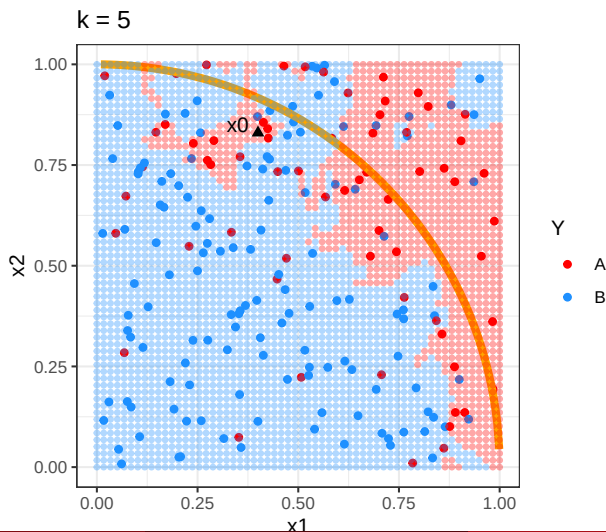
$k=2$

Here are the classification boundaries for a variety of values of K .



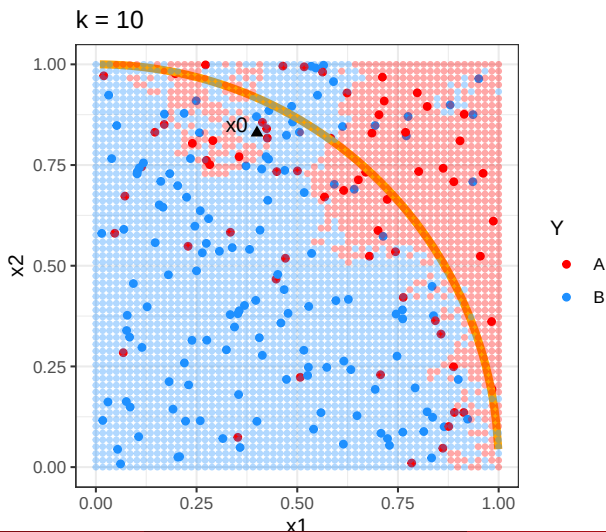
$k=5$

Here are the classification boundaries for a variety of values of K .



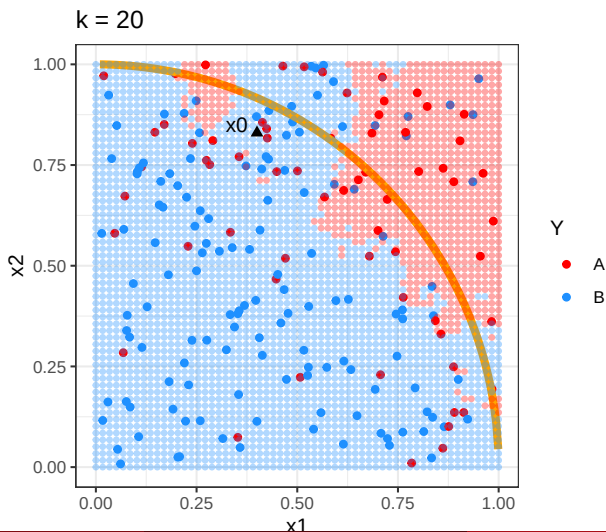
$k=10$

Here are the classification boundaries for a variety of values of K .



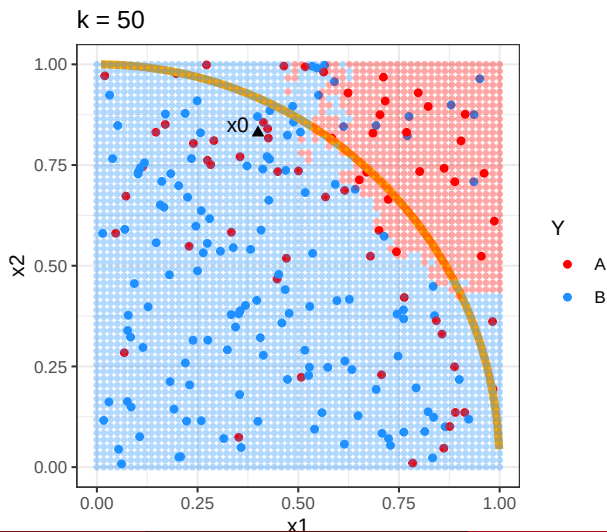
$k=20$

Here are the classification boundaries for a variety of values of K .



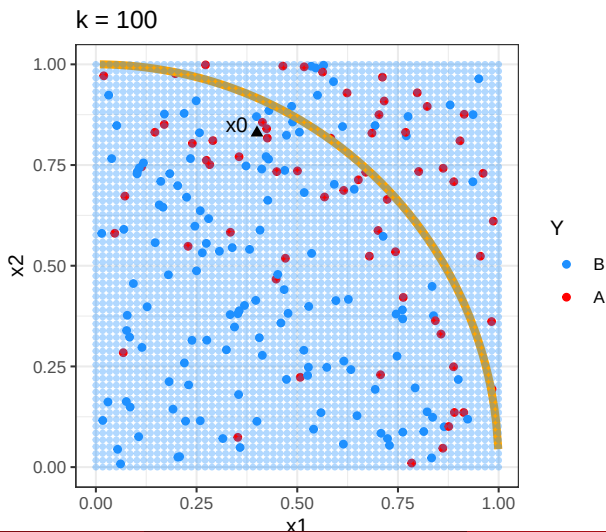
$k=50$

Here are the classification boundaries for a variety of values of K .



$k=100$

Here are the classification boundaries for a variety of values of K .



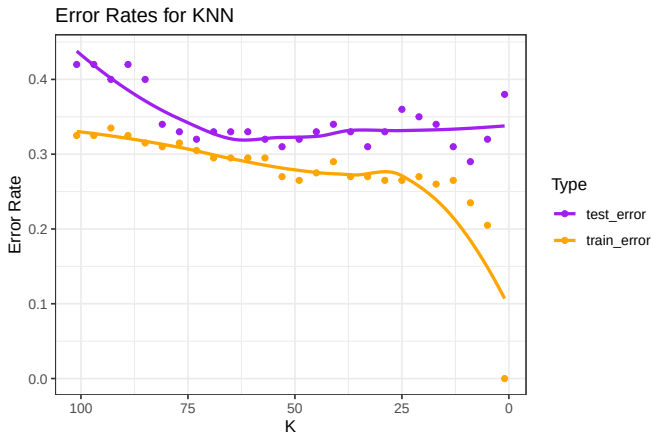
Questions

Q: Which seemed like the “best” K ?

Q: As K increased, did the classification boundaries get simpler, or more complex?

Error Rates

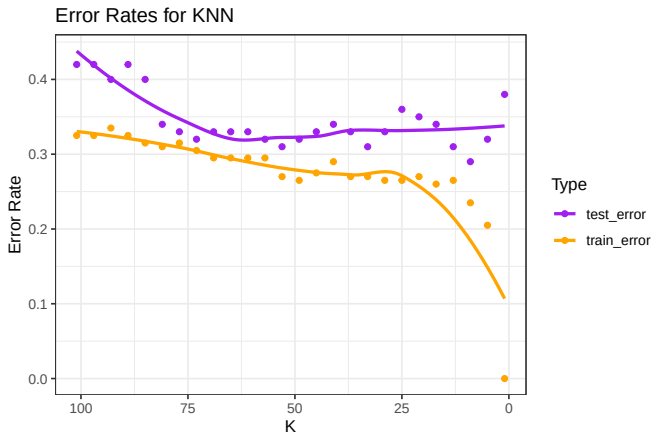
The graph below shows error rates for the training set, as well as a test set of 100 points.



Q: What does this look similar to?

Error Rates

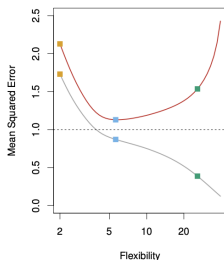
The graph below shows error rates for the training set, as well as a test set of 100 points.



Q: What does this look similar to? **Model complexity U curve!**

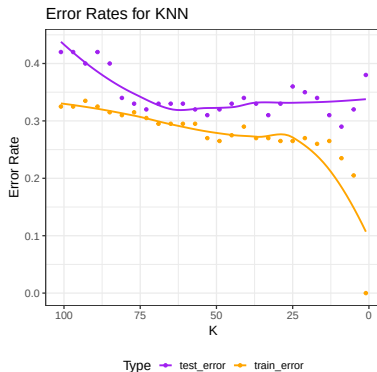
Model complexity and K in KNN

Figure 2.9 from *ISLR*



- RED: Test MSE
- GRAY: Training MSE
- X-axis is model complexity

KNN Error Rates



Moral: The K in KNN is *inversely* related to model complexity

Note: KNN for Quantitative Response Variables

KNN is very similar when predicting a **quantitative** response, Y

Note: KNN for Quantitative Response Variables

KNN is very similar when predicting a **quantitative** response, Y

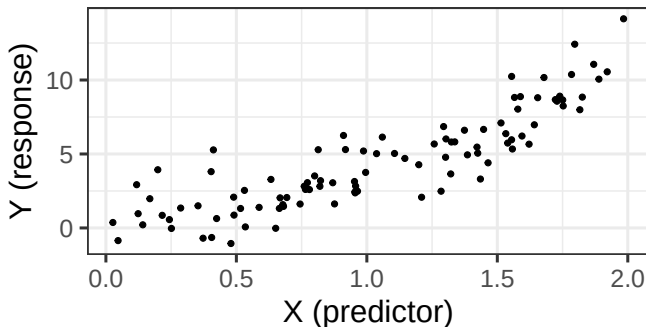
- Assign x_0 the **average** response among its K nearest neighbors

Note: KNN for Quantitative Response Variables

KNN is very similar when predicting a **quantitative** response, Y

- Assign x_0 the **average** response among its K nearest neighbors

Training Data

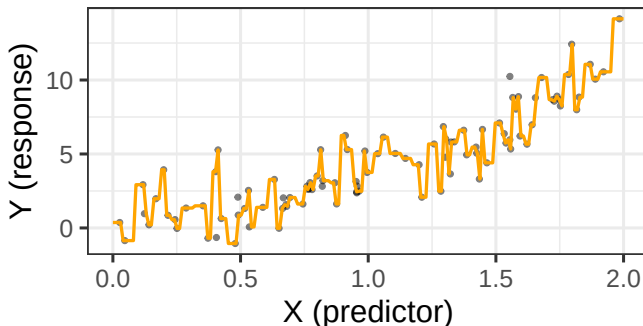


Note: KNN for Quantitative Response Variables

KNN is very similar when predicting a **quantitative** response, Y

- Assign x_0 the **average** response among its K nearest neighbors

KNN (K=1)

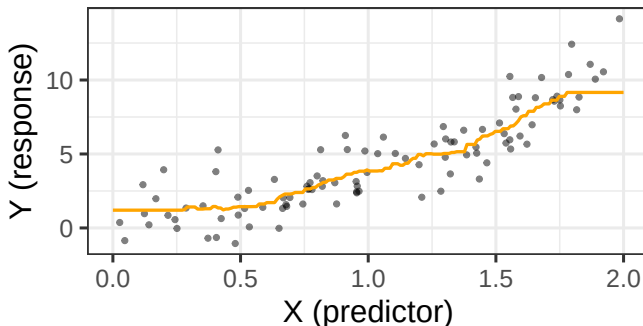


Note: KNN for Quantitative Response Variables

KNN is very similar when predicting a **quantitative** response, Y

- Assign x_0 the **average** response among its K nearest neighbors

KNN (K=20)

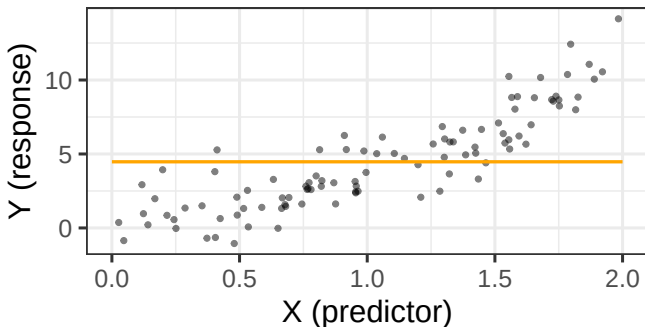


Note: KNN for Quantitative Response Variables

KNN is very similar when predicting a **quantitative** response, Y

- Assign x_0 the **average** response among its K nearest neighbors

KNN (K=100)



Section 3

Recap

Recap

Today, we discussed

- The theoretical best classification model, the **Bayes Classifier**
- **K-Nearest Neighbors (KNN)** for Classification (and regression!)

Rest of Class:

- Practice KNN on HW 3, Problem 6